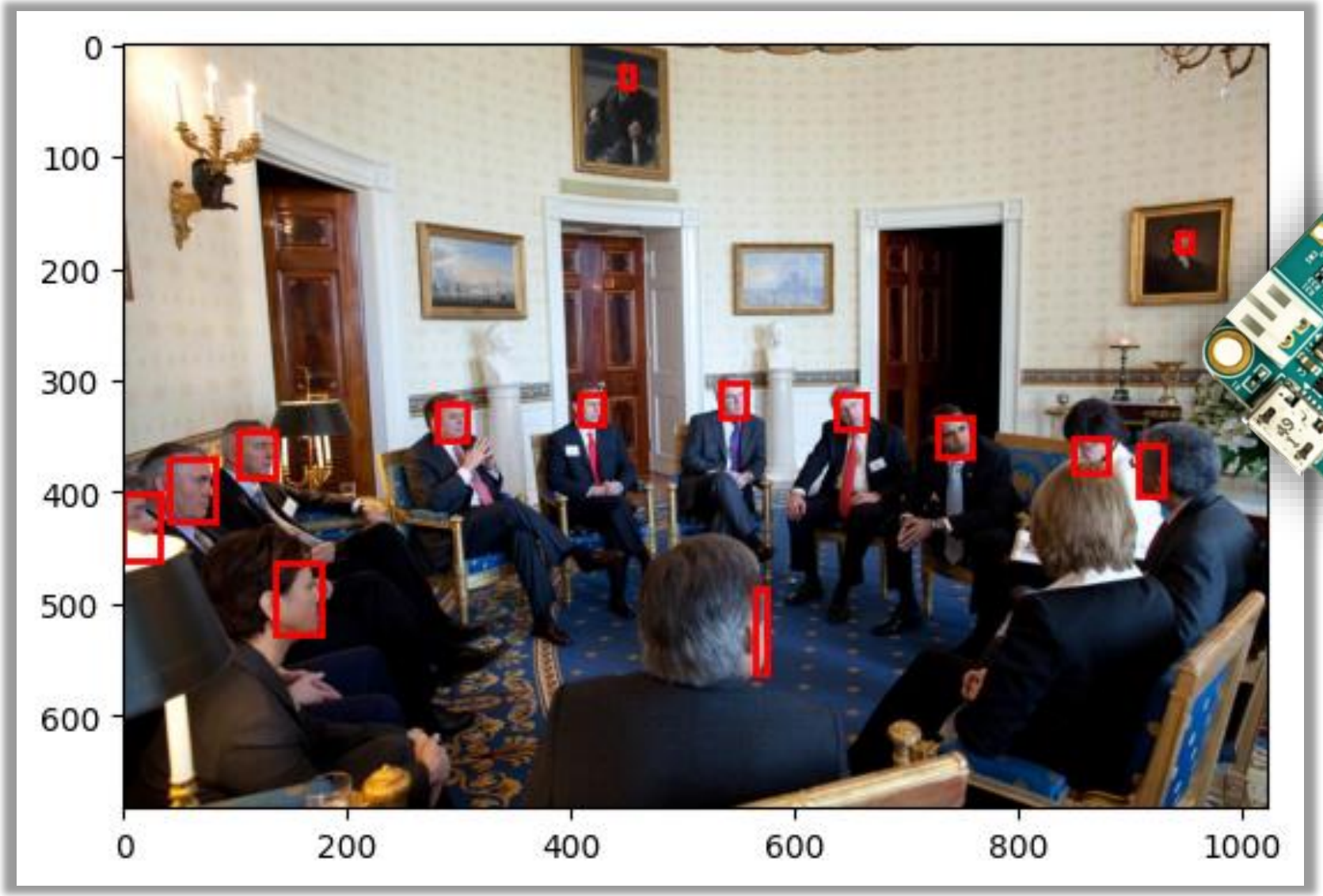
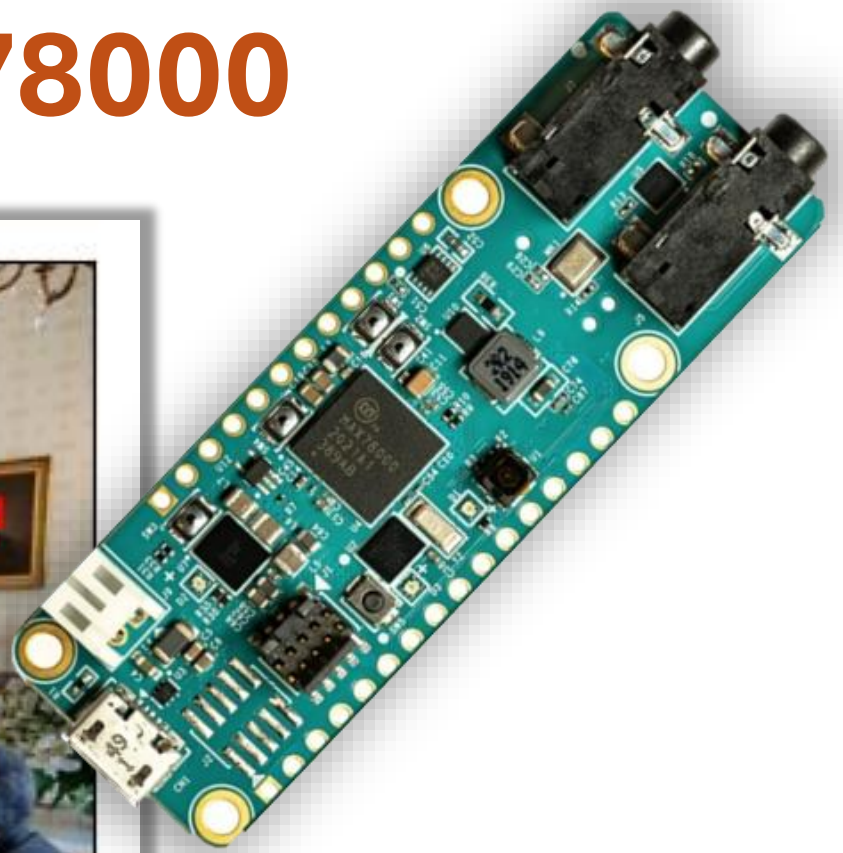
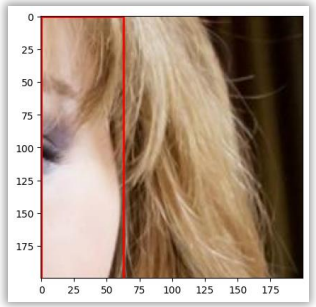
Face detection using MAX78000

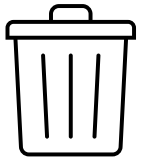




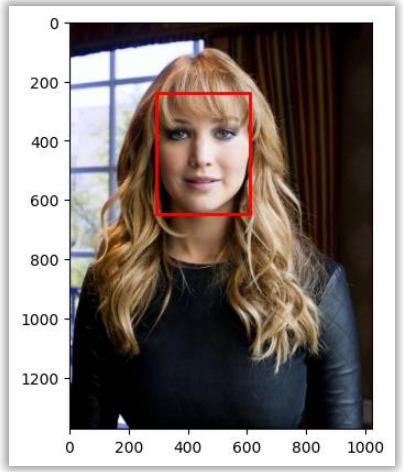
Dataset preprocessing

Random 200×200 cropping

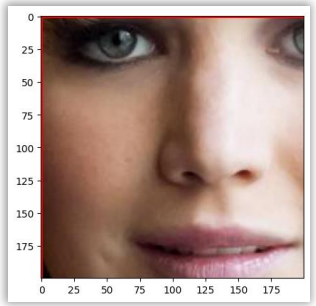




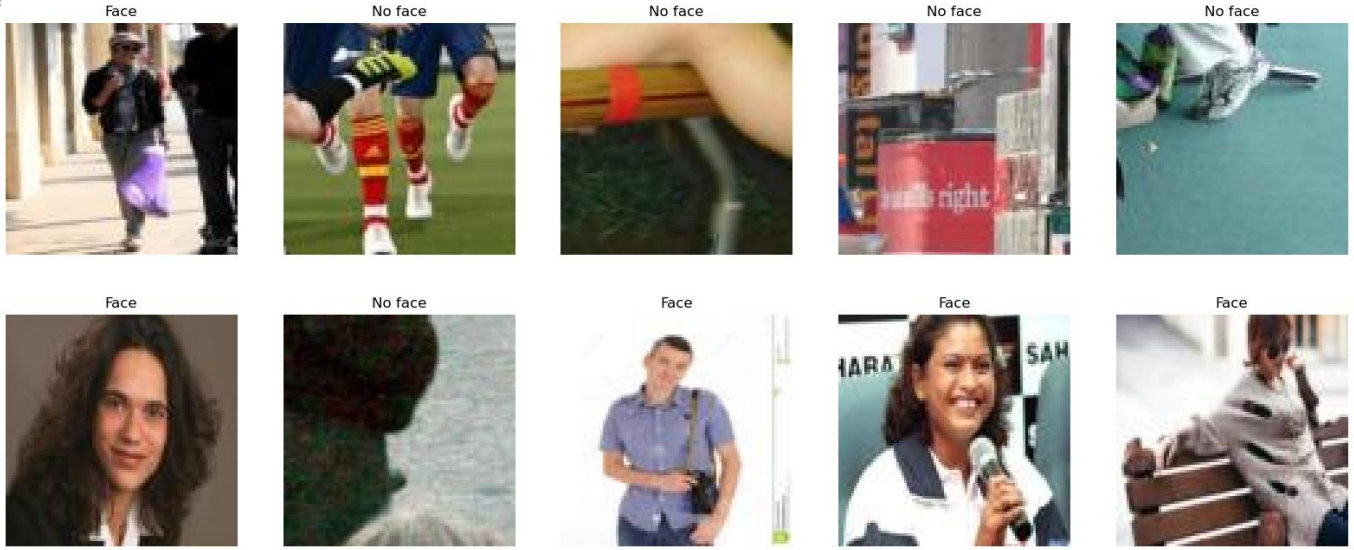
Removing ambiguous images



WIDERface



88×88 resize



Classification

Training

Training with early stopping

Initial Training with Early Stopping :

- Trained the network until the validation loss stopped improving.
- Determined the best epoch (E).

QAT with reduced LR

Quantization-Aware Training :

- Activated QAT at epoch E.
- Divided the learning rate by 10 starting from epoch E.
- Continued training with early stopping to determine the best QAT network (N)

Post-QAT consistency check

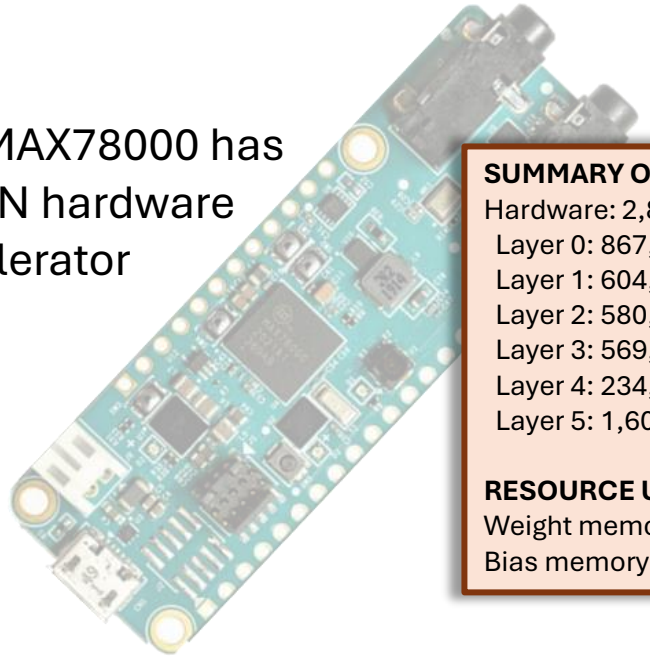
Quantization Verification :

- Quantized network N and verified that it achieved results consistent with those during training.

Optimizer	Adam
Learning rate	0,001
Weight decay	0,0001
Momentum	0,9
Dampening	0

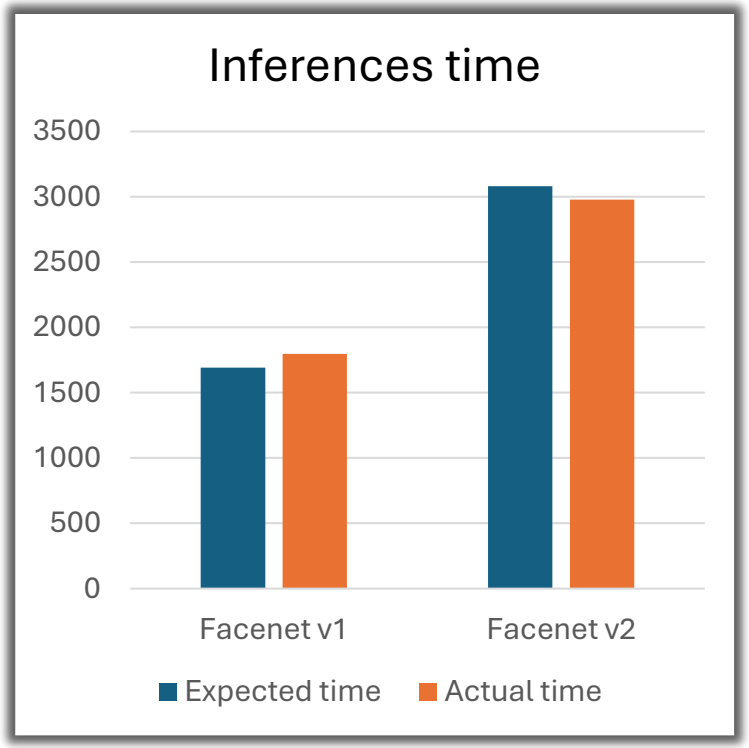
Predicting inference time

The MAX78000 has a CNN hardware accelerator

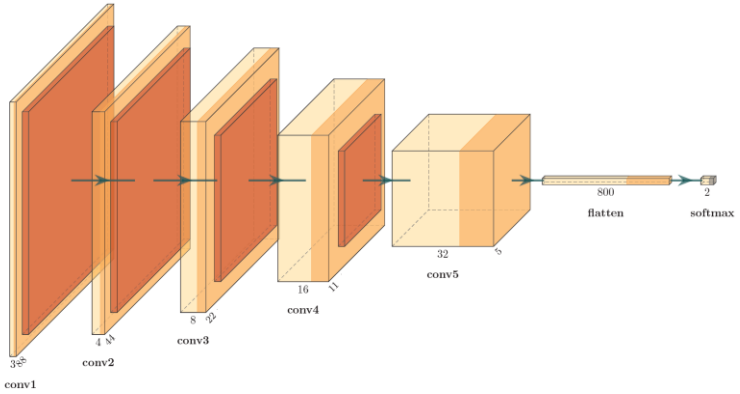


SUMMARY OF OPS
Hardware: 2,857,344 ops (2,741,056 macc; 116,288 comp; 0 add; 0 mul; 0 bitwise)
Layer 0: 867,328 ops (836,352 macc; 30,976 comp; 0 add; 0 mul; 0 bitwise)
Layer 1: 604,032 ops (557,568 macc; 46,464 comp; 0 add; 0 mul; 0 bitwise)
Layer 2: 580,800 ops (557,568 macc; 23,232 comp; 0 add; 0 mul; 0 bitwise)
Layer 3: 569,184 ops (557,568 macc; 11,616 comp; 0 add; 0 mul; 0 bitwise)
Layer 4: 234,400 ops (230,400 macc; 4,000 comp; 0 add; 0 mul; 0 bitwise)
Layer 5: 1,600 ops (1,600 macc; 0 comp; 0 add; 0 mul; 0 bitwise)

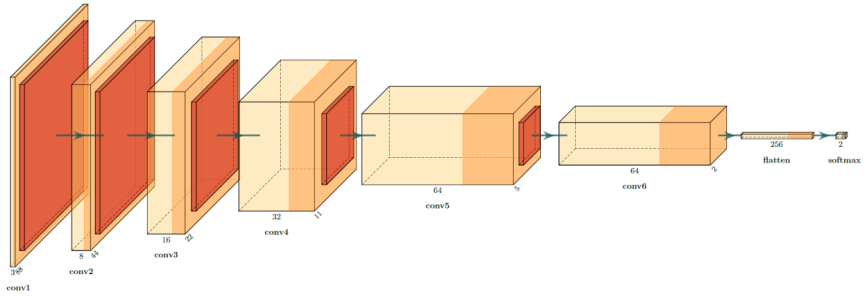
RESOURCE USAGE
Weight memory: 16,972 bytes out of 442,368 bytes total (3.8%)
Bias memory: 2 bytes out of 2,048 bytes total (0.1%)



Hardware accelerator
64 processors in parallel
Running at 50MZ
Dealing with 3×3 kernels



Facenet v1, accuracy 85%



Facenet v2, accuracy 86%

Further quantization

SUMMARY OF OPS
Hardware: 9,666,944 ops (9,433,088 macc; 233,856 comp; 0 add; 0 mul; 0 bitwise)
Layer 0: 1,734,656 ops (1,672,704 macc; 61,952 comp; 0 add; 0 mul; 0 bitwise)
Layer 1: 2,323,200 ops (2,230,272 macc; 92,928 comp; 0 add; 0 mul; 0 bitwise)
Layer 2: 2,276,736 ops (2,230,272 macc; 46,464 comp; 0 add; 0 mul; 0 bitwise)
Layer 3: 2,253,504 ops (2,230,272 macc; 23,232 comp; 0 add; 0 mul; 0 bitwise)
Layer 4: 929,600 ops (921,600 macc; 8,000 comp; 0 add; 0 mul; 0 bitwise)
Layer 5: 148,736 ops (147,456 macc; 1,280 comp; 0 add; 0 mul; 0 bitwise)
Layer 6: 512 ops (512 macc; 0 comp; 0 add; 0 mul; 0 bitwise)

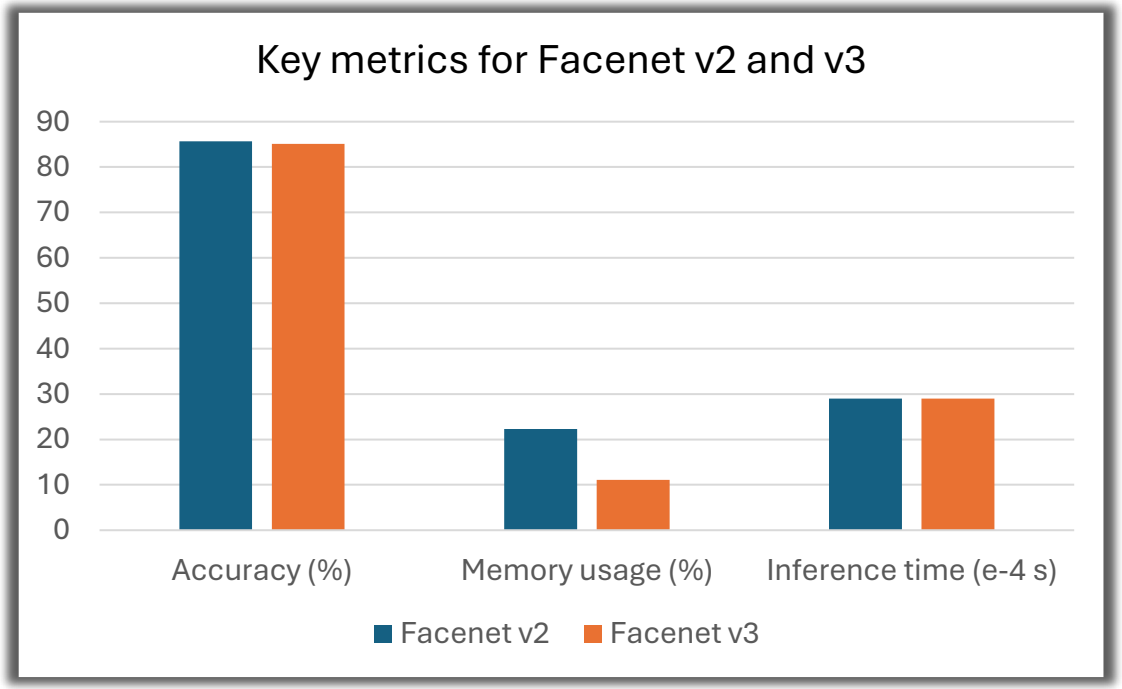
RESOURCE USAGE
Weight memory: 49,324 bytes out of 442,368 bytes total (11.1%)
Bias memory: 2 bytes out of 2,048 bytes total (0.1%)

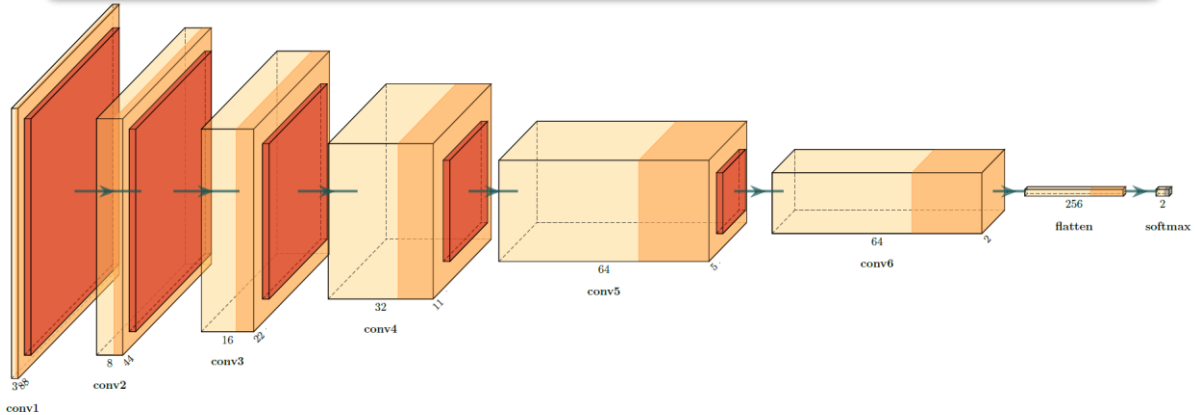
Facenet v3

SUMMARY OF OPS
Hardware: 9,666,944 ops (9,433,088 macc; 233,856 comp; 0 add; 0 mul; 0 bitwise)
Layer 0: 1,734,656 ops (1,672,704 macc; 61,952 comp; 0 add; 0 mul; 0 bitwise)
Layer 1: 2,323,200 ops (2,230,272 macc; 92,928 comp; 0 add; 0 mul; 0 bitwise)
Layer 2: 2,276,736 ops (2,230,272 macc; 46,464 comp; 0 add; 0 mul; 0 bitwise)
Layer 3: 2,253,504 ops (2,230,272 macc; 23,232 comp; 0 add; 0 mul; 0 bitwise)
Layer 4: 929,600 ops (921,600 macc; 8,000 comp; 0 add; 0 mul; 0 bitwise)
Layer 5: 148,736 ops (147,456 macc; 1,280 comp; 0 add; 0 mul; 0 bitwise)
Layer 6: 512 ops (512 macc; 0 comp; 0 add; 0 mul; 0 bitwise)

RESOURCE USAGE
Weight memory: 98,648 bytes out of 442,368 bytes total (22.3%)
Bias memory: 2 bytes out of 2,048 bytes total (0.1%)

Facenet v2





Facenet v2 and v3 architecture

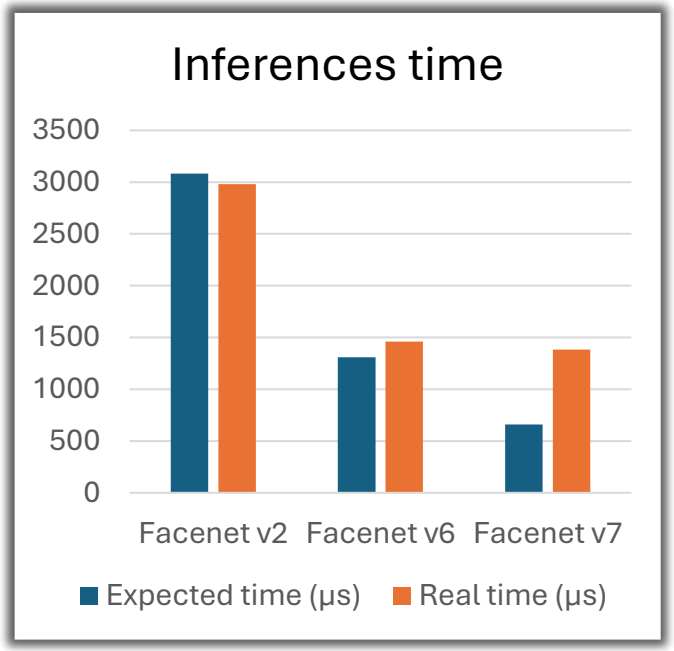
Optimizing inference time

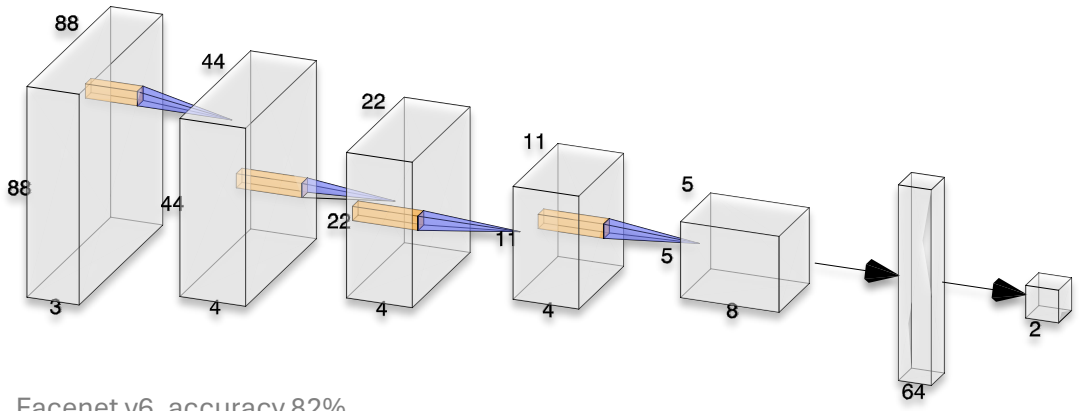
Facenet v2	
Layer 1	1652.05 µs
Layer 2	851.84 µs
Layer 3	367.84 µs
Layer 4	169.40 µs
Layer 5	34.50 µs
Layer 6	5.52 µs
Layer 7	0.16 µs
TOTAL	**3081.31 µs**

Facenet v6	
Layer 1	826.03 µs
Layer 2	348.48 µs
Layer 3	87.12 µs
Layer 4	33.88 µs
Layer 5	11.00 µs
Layer 6	1.00 µs
TOTAL	**1307.51 µs**

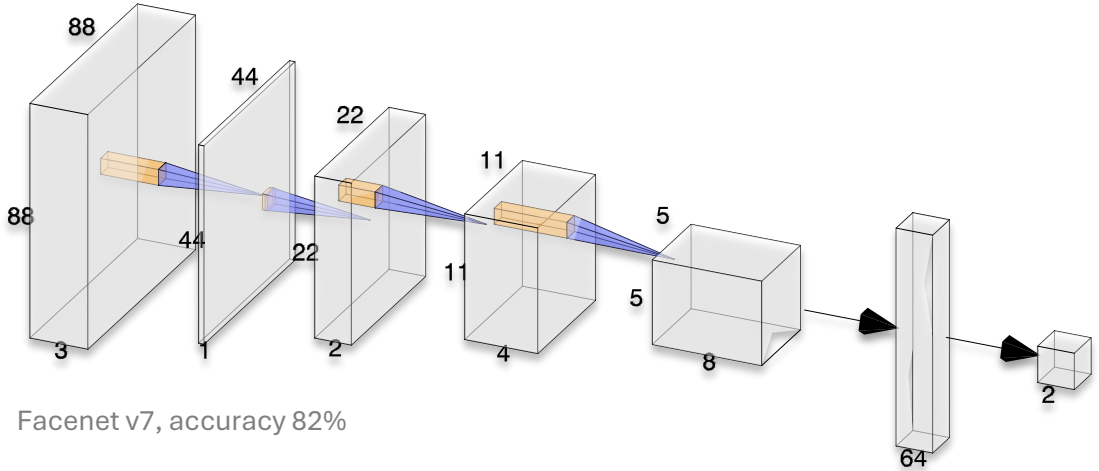
Facenet v7	
Layer 1	206.51 µs
Layer 2	309.76 µs
Layer 3	96.80 µs
Layer 4	33.88 µs
Layer 5	11.00 µs
Layer 6	1.00 µs
TOTAL	**658.95 µs**

Expected inference time per layer



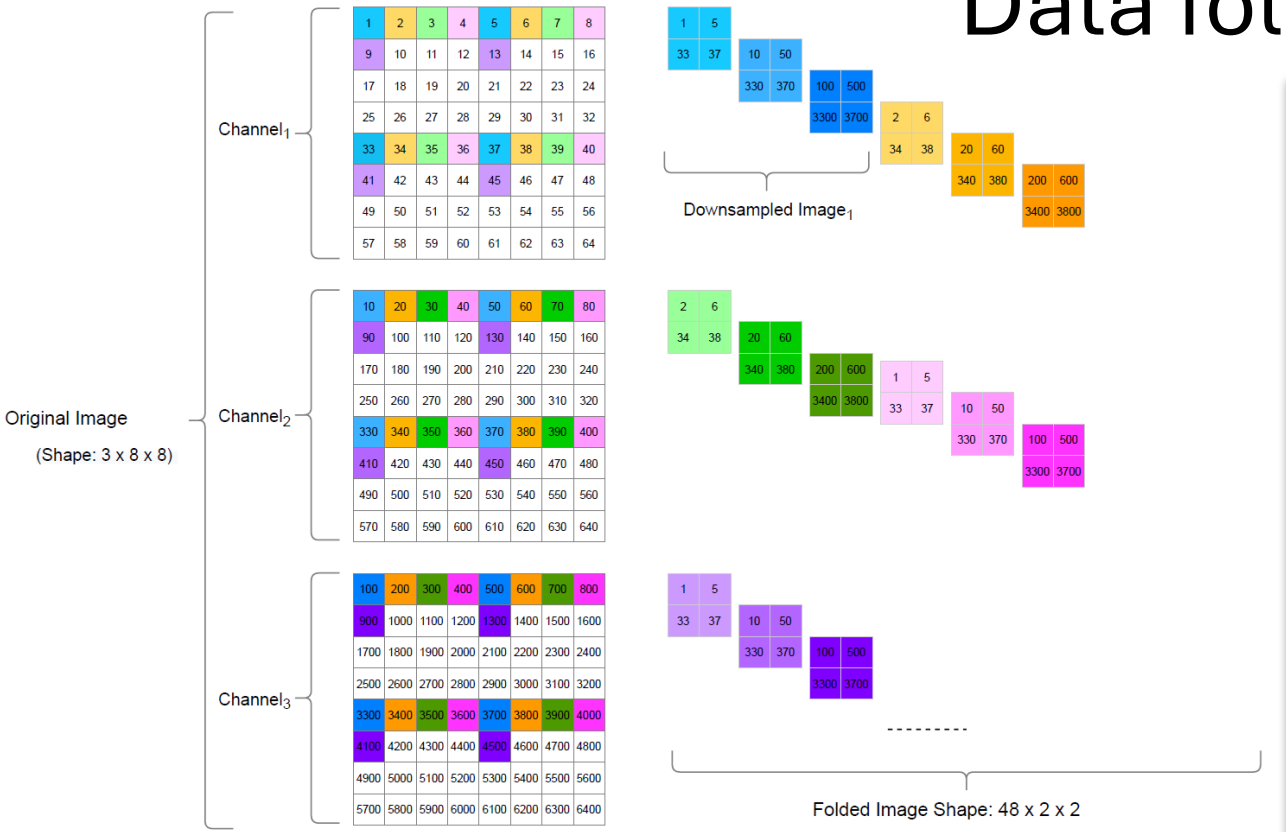


Facenet v6, accuracy 82%



Facenet v7, accuracy 82%

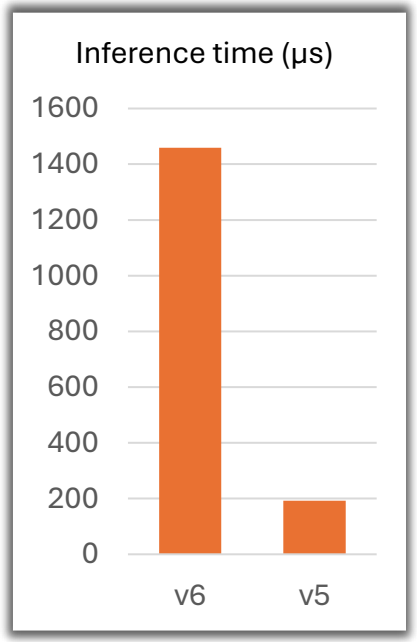
Data folding

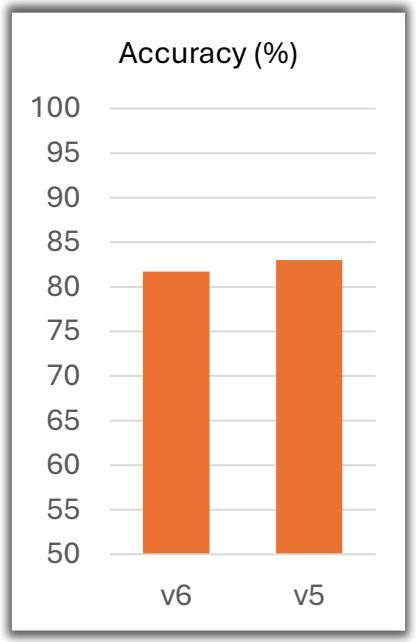


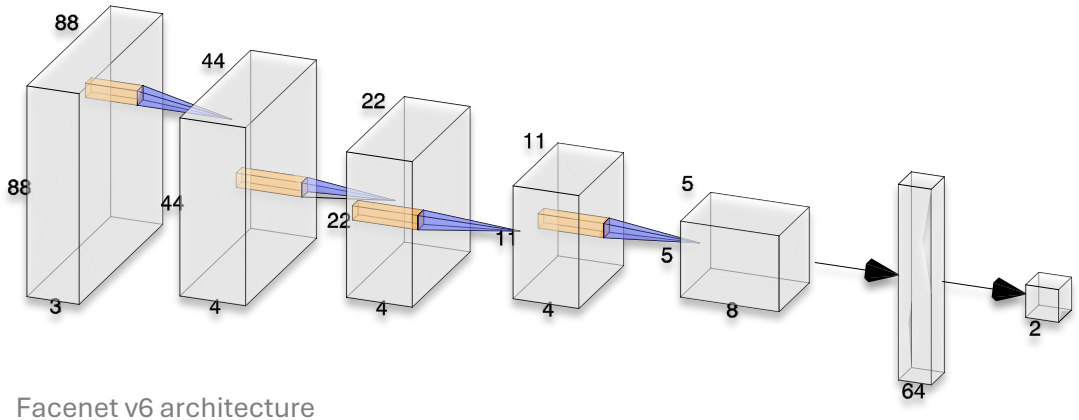
Data folding according to L^3U-net by Okman & al. [1]

Facenet v6	
Layer 1	826.03 µs
Layer 2	348.48 µs
Layer 3	87.12 µs
Layer 4	33.88 µs
Layer 5	11.00 µs
Layer 6	1.00 µs
TOTAL	**1307.5 µs**

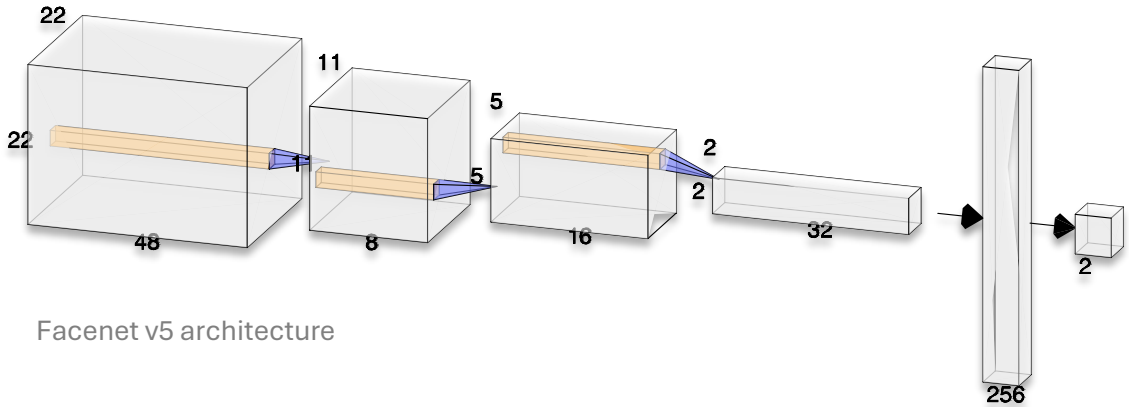
Facenet v5	
Layer 1	79.05 µs
Layer 2	53.24 µs
Layer 3	19.00 µs
Layer 4	5.60 µs
Layer 5	0.16 µs
TOTAL	**157.05 µs**







Facenet v6 architecture



Facenet v5 architecture

References

- Okman, Ulkar & Uyanik, 2022. L3U-net: Low-Latency Lightweight U-net Based Image Segmentation Model for Parallel CNN Processor, http://arxiv.org/abs/2203.16528 [1]